# Defining an Optimal Metric for the Path Collective Variables

Ladislav Hovan,[†] Federico Comitani,[†] and Francesco L. Gervasio[*,†,‡]

†*Department of Chemistry, University College London, London WC1E 6BT, United Kingdom*

‡*Institute of Structural and Molecular Biology, University College London, London WC1E 6BT, United Kingdom*

E-mail: f.l.gervasio@ucl.ac.uk

**Abstract**

Path Collective Variables (PCVs) are a set of path-like variables that have been successfully used to investigate complex chemical and biological processes and compute their associated free energy surfaces and kinetics. Their current implementation relies on general, but at times inefficient, metrics (such as RMSD or DRMSD) to evaluate the distance between the instantaneous conformational state during the simulation and the reference coordinates defining the path. In this work, we present a new algorithm to construct optimal PCVs metrics as linear combinations of different CVs weighted through a spectral gap optimization procedure. The method was tested first on a simple model, trialanine peptide *in vacuo* and then on a more complex path of an anticancer inhibitor binding to its pharmacological target. We also compared the results to those obtained with other path-based algorithms. We find that not only our proposed approach is able to automatically select relevant CVs for the PCVs metric, but also that the resulting PCVs allow to reconstruct the associated free energy very

efficiently. What is more, at difference with other path-based methods, our algorithm is able to explore non-locally the reaction path space.

# 1 Introduction

Atomistic molecular dynamics simulations (MD) are a powerful technique whose success in investigating complex physical, chemical, and biological systems increased in parallel with advances in computing power[1–4] and force field accuracy.[5–7] Many events of interest, however, have characteristic timescales that are much longer than those accessible by unbiased MD, even when running on specialized supercomputers. To address this limitation, over the years a number of algorithms have been developed to enhance the sampling and reconstruct the associated free energy landscape.[8,9] Two families of algorithms have been particularly successful: one is based on the definition of a path connecting the two end states of a biological process of interest (e.g. bound and unbound states of a ligand to its target);[10,11] while the other relies on a set of explicit descriptors that approximate the reaction coordinate, known as collective variables (or CVs).[12–17]

Path-based methods, like Transition Path Sampling,[10] Discrete Path Sampling,[18] Milestoning,[11] Nudged Elastic Band[19] and the Finite Temperature String method,[20] require some knowledge of the endpoint states and often also an initial guess path. Methods based on CVs, like Umbrella Sampling[13] or Metadynamics,[14,21,22] can instead evolve spontaneously to the end state(s) when an optimal set of CVs approximating the reaction coordinate is provided. The Path Collective Variables (PCVs)[23] combine various aspects of both these approaches; they describe the progression along (s) and the distance from (z) an initial (guess) path in the free energy space, by means of two functions mathematically defined as:

$$s = \frac{\sum_{i=1}^{N} i \exp(-\lambda R[X - X_i])}{\sum_{i=1}^{N} \exp(-\lambda R[X - X_i])}, \tag{1}$$

$$z = -\frac{1}{\lambda} \ln \left[ \sum_{i=1}^{N} \exp(-\lambda R[X - X_i]) \right]. \quad (2)$$

In the original implementation, the path was defined in terms of a series of sequential structures (snapshots) of the system under investigation describing the transition of interest and extracted from preliminary simulations. In the above equations, $X$ represents the atomic coordinates at the current simulation time-step, while $X_i$ denotes those of the i-th snapshot. The function $R$ represents here a chosen metric, which measures the distance between configuration states. The $\lambda$ parameter serves to smooth the variation of the $s$ variable. With this definition, $s$ takes values between 1 and $N$, where $N$ is the total number of snapshots, but it can be easily normalized to a $\langle 0, 1 \rangle$ range. The range of $z$ depends on the choice of the metric, but it can easily be seen that its value falls to 0 when the system exactly matches a snapshot, while it increases monotonically as the system moves away from the reference path.

PCVs can be used with well-tempered Metadynamics (MetaD)[14,24] to enhance the sampling and allow for a non-local exploration of the free energy surface, including optimal (low free energy) paths that might be far from the initial guess path and inaccessible to many other path-based approaches. In this sense, PCVs and the more recent Path-Metadynamics[25] combine the advantages of path-based methods with those of CVs. The knowledge of end states is still required, but for many interesting systems they are known, as in the case of the binding of ligands to their pharmacological targets.

Indeed path-based methodologies are inherently suitable for investigating ligand binding events as these can be naturally described by paths connecting the bound and a set of unbound states. Metadynamics with PCVs has been successfully employed to determine the free energy landscape of binding in a number of interesting cases, such as protein kinases CDK2,[26] p38[27] and c-Src Kinase,[28] the $\beta$2-adrenergic receptor,[29] and cyclooxygenase enzymes COX1 and COX2.[30,31] Often, when a family of compounds has to be tested against a single target protein, the same path can be used as a reference for all ligands, allowing for

an efficient comparison of binding free energies.[26] In this context, PCVs outperforms many other geometry-based CVs when used with Metadynamics to predict the binding energies, as they focus the exploration on relevant regions of the free energy landscape.[26,27] In the past, our group has also combined PCVs with a transition-path-sampling approach, partial path transition interface sampling (PPTIS), to obtain a unified framework (TS-PPTIS) and compute both the kinetics and the thermodynamics associated with complex events such as protein folding and ligand binding.[32]

As discussed above, the definition of PCVs requires a metric to quantify the progression along the path and the distance from it, for instance by measuring the instantaneous difference from reference snapshots. In the original implementation, the Root Mean Square Deviation (RMSD) was the metric of choice, requiring the alignment of structures to the reference, mainly performed using the Kearsley algorithm.[33] Since several structures need to be aligned at each time-step (one for each reference snapshot), the computational cost of this approach can be significant and the risk of misaligning structures makes it unreliable. A more serious limitation of an RMSD metric stems from its difficulty in distinguishing conformations that are similarly "faraway" or dissimilar from the reference structure. This makes the exploration of low free energy regions far from the initial guess path more problematic. As an alternative to RMSD, the distance-RMSD, or DRMSD, which measures the differences between atomic distances within structures, or a contact map matrix[34] were also implemented in the popular plug-in for free energy calculations PLUMED.[35] These metrics do not require alignment but are also affected by the inability to fully resolve the conformational degeneracy arising when leaving the reference snapshot. Moreover, the corresponding $z$ variables are of difficult interpretation, at variance from RMSD, where the distance from the snapshots is in units of length (typically Angstroms). Other metrics, such as one based on chirality,[36] are more situational and suitable for specific systems.

For a number of years, it has been clear that redefining the PCVs with a metric that itself combines a number of different CVs would be highly desirable, as it would allow to directly

capture more complex motions and possibly solve the degeneracy problem at high $z$. However, it is still not clear how such CVs could be chosen and how to identify the correct weights.

In this work, we present an approach inspired by the spectral gap optimization (or SGOOP) recently proposed by Tiwary et al.,[37] that allows to define the metric as a linear combination of CVs selected from a pool of possible variables. We have called this method COMet-Path (Coefficients Optimization of a Metric for Path Collective Variables).

SGOOP allows to identify an optimal set of CVs by selecting those maximizing the spectral gap between fast and slow eigenvalues in the transition probability matrix. The spectral gap contains information on the timescale separation between fast and slow dynamics and allows the identification of variables that drive the biologically relevant dynamics.[38,39] For a more detailed description see Ref. 37. Here, we build on this approach, but instead of trying to explore the space of possible CVs directly, we adopt the Path-CVs and use the spectral gap maximization to optimize their metric.

We start from a metric comprised of a set of simple collective variables $x_j$ with different coefficients (or weights) $c_j$, defined as:

$$R[X - X_i] = \sum_{j=1}^{M} c_j^2 (x_j - x_{i,j})^2. \tag{3}$$

The set can include a large and differentiated pool of basic CVs relevant to the process under study (that might be distances, contact maps, angles, etc...).

Given an initial, potentially large, set of variables, one needs to determine which ones to use in the final PCV metric and what are their optimal coefficients. A way of doing this was presented by Dixit et al.[40] and is described in detail in the Supporting Information. Briefly, this approach allows us to relate the transition rates between basins to their stationary populations. From these rates, the degree of separation between the slow and fast degrees of freedom can be estimated, and the best set of variables can thus be identified.

We start from a discretized Markovian random walker on a directed network with nodes

$V = \{a, b, \cdots\}$ and edges $E$, and assume the probabilities are normalized and stationary. One can then obtain the transition rate between two edge node states, $a$ and $b$ of the network, as:

$$\omega_{ab} = \mu\sqrt{\frac{p_b}{p_a}}, \qquad \text{if}(a,b) \in E. \tag{4}$$

$p_a$ and $p_b$ are the probabilities of visiting the two states and the parameter $\mu$ is related to the Lagrange multiplier for the normalization condition $\gamma$ via the following relation: $\mu\delta t = e^{-\gamma}$. The values of $\omega_{ab}$ are the elements of the transition matrix $\Omega$.

Now consider the ordered eigenvalues $\lambda$ of the transition probability matrix $\mathbf{\Omega}$: $\lambda_0 \equiv 1 > \lambda_1 \geq \lambda_2$.... If there are $s$ barriers apparent from the free energy projection estimate reweighted as a function of the chosen CVs and their weights,[41] the spectral gap is simply defined as $\lambda_s - \lambda_{s+1}$. The slowest eigenvalues correspond to the slowest changes in the system and represent the transition over the main barriers in the free energy surface. By maximizing the spectral gap we are thus increasing the separation between these and the fast eigenvalues. The energy threshold for the barriers can be tuned by the user. This choice is responsible for the separation between slow and fast processes, so it's sensible to keep it in the order of $k_BT$, since barriers less then this threshold can be easily crossed by an unbiased system at room temperature, categorizing them as fast processes.

The derivation of Eq. 4 assumes equilibrium between all the nodes of the network. It also uses only the first two terms of the Taylor expansion of the transition matrix $\mathbf{k}$ , which is valid for small values of $\Delta t$. We also assume that there are no jumps in the transitions of the s variable, i.e. that the rate of the transitions between non-adjacent states is zero.

Our method employs an iterative optimization procedure, as illustrated schematically in Fig. 1: first, a trial Metadynamics run is performed with a simple combination of collective variables in order to get a crude estimate of the stationary probabilities in each basin; then,

a subset of CVs to be optimized is chosen. A large set of sequential frames is generated from the trajectory, typically starting from a small group of chosen "landmark" states. Then, the space of the coefficients is explored through cycles of simulated annealing, starting from equal initial values. At each step, a series of frames is chosen from the larger set describing the path, while imposing equal spacing between them. The perturbations made to the coefficients are small, and the negative of the spectral gap in the path collective variable $s$ is used as an energy function. The convergence of the simulated annealing process is determined by tracking the best spectral gap. The cycles are then terminated when this reaches a set threshold obtained by counting the number of barriers in the free energy surface which are higher than an adjustable value (around 1 kT).

The computational cost of the optimization process depends on the extent of the initial simulation and the number of CVs under consideration. In our case, using a sequential C++ code running on a Intel® Core™ i5-5200U processor, we were able to test a single combination of variables on the results of a 200 ns initial simulation in less than 1s. Screening tens of thousands of combination might take several hours on a similar hardware, making it orders of magnitude less computationally demanding than typical MD simulations on complex systems.

An added bonus of our approach is that one could use different coefficients in different sections of the path where the corresponding slow variables might differ.[42,43]

Here, we tested our method against two different exemplary problems: first to sample the free energy landscape of the trialanine peptide as a function of a path defined by its dihedral angles, and second to characterize the binding of Dasatinib to the c-Src kinase; the former was chosen for its simplicity and its free energy landscape, whose shape is well known, while the latter allowed us to test the capabilities of our method against a known real case scenario.

## 2 Results and Discussion

**Trialanine**. As a first simple test case, trialanine (Ace-$Ala_3$-Nme) *in vacuo* was investigated. The molecule is typically chosen to test free energy methods for its small size, thus requiring less computational resources, and its characteristic free energy surface (FES), which has several minima separated by relatively high barriers. Trialanine represents a substantially more complex test case than the commonly employed alanine dipeptide and it was also chosen as a test system for SGOOP,[37] making it ideal for a direct comparison.

Trialanine has six backbone dihedral angles (three $\Phi$ and three $\Psi$). In Ref. 37, it was suggested that the $\Phi$ angles have more impact in the free energy shape. Therefore, we run a preliminary MetaD to reconstruct the FES as a function of these three variables. In Fig. 3A the projection onto the first two angles (here labeled $\Phi_1$ and $\Phi_2$), and the paths used to test our method are shown. These paths were chosen in such a way that the third $\Phi$ angle, not shown in the figure, could be neglected with no major consequences. Only one angle at a time ($\Phi_1$ and $\Phi_2$, respectively) are needed to move along paths 1 and 2, while both of them are needed for a proper description of paths 3 and 4.

The paths were built automatically through a Monte Carlo procedure that first randomly selects a group of reference frames along the path and then iteratively optimizes this choice to assure uniformity in the distance between neighboring frames.

The results of our COMet-Path algorithm are shown in Fig. 3 B. For each path, the highest coefficients (highlighted in bold) correspond to the most important variables for the given transition. As expected, the algorithm is able to correctly identify which angles are crucial for each respective path, $\Phi_1$ for path 1, which is weighted at 0.82, $\Phi_2$ in path 2, weighted 0.77 and both $Phi_1$ and $Phi_2$ with approximatively equal weights in the last two cases. Relatively little weight is assigned to the unimportant angle $\Phi_3$.

It also is interesting to note that while in the case of the first three paths, the number of barriers identified corresponds to the number of free energy barriers on the initial path, this is not the case for the fourth path. This last path was intentionally chosen to be far from

ideal and passing through a local maximum, however, the reweighting procedure was able to correctly count the two barriers on the lowest energy path connecting A to B.

Since the first two paths refer to the same system, use the same input data and have the same endpoints, their spectral gaps values are directly comparable. Both are lower than that of the third path, and so is their separation of the slow and fast degrees of freedom. Similarly, the fourth path has a lower spectral gap when compared to the third path, suggesting that the latter would perform better.

To compare the performance of COMet-Path to that of PCVs with an RMSD-based metric, two simulations (with the two metrics) were run on snapshots taken from path 3. The results can be seen in Fig. 4. The free energy surfaces reweighted as a function of the first two $\phi$ angles are qualitatively similar, but COMet-Path (panel A) samples more the relevant states, including the region of the transition states, resulting overall in a much better defined free energy landscape. The evolution of the 1D free energy estimates along the $s$ variable is also shown for the three minima. The estimates converge very quickly, within 100ns, to the reference values when using COMet-Path (panel C) while RMSD estimates for the minimum C fail to properly converge to the expected value even after twice the same simulation time (panel D).

**Dasatinib binding to c-Src**. As a second, more complex and realistic test case, we investigated the binding of an anticancer inhibitor Dasatinib to the c-Src kinase. This system has been previously studied in our group by means of PCVs with an RMSD metric and in combination with parallel tempering.[28] The problem of determining the binding energy of ligands where the binding pose is known is indeed well suited for path-based methods. The initial path estimate can be obtained from a short biased unbinding simulation which doesn't necessarily need to converge the free energy. In the case of COMet-Path selected PCV, the data from this simulation can then be used also during the optimization stage.

In this case, we have chosen a small set of possible variables which we believed to be relevant to the problem; given their simplicity, they wouldn't be able to describe the transition

when taken alone. The variables along with their optimized coefficients values can be seen in Table 2 in Supplementary Information.

We have then performed a 1 $\mu s$ MD simulation using COMet-Path with weak harmonic restraints on the $z$ variable and without volume restraints or parallel tempering. Interestingly, during the Metadynamics run, several recrossing were observed, which normally would have required multiple replicas in a combined parallel tempering Metadynamics simulation to be used when employing simple geometric variables.

The resulting free energy surface is shown in Fig. 3 C. As expected, the range of values explored along the $z$ variable is limited in the bound state, while in the unbound state the system can easily visit configurations far from the reference snapshots. The exploration of the unbound states could be limited with stricter restraints to speed up the convergence of the free energy, but we chose here not to apply any external bias.

In this case, the values at high $s$ correspond to the unbound state and we had little interest in distinguishing them, however, it is possible to include more variables to detect either internal changes within Dasatinib or the interaction with protein surface regions outside the binding site. The final free energy surface obtained with COMet-Path is shown in Fig. 3 C, while details on the convergence of this simulation are found in the SI.

To evaluate the accuracy of our results, we can compare the free energy profile along the $s$ variable for the COMet-Path and the equivalent on the RMSD Path (in Fig. 3 D). It is important to clarify that similar values of $s$ in the two profiles do not necessarily correspond to the same conformation, given that the snapshots were chosen in such a way to be equally spaced in different metrics. However, the path explored is the same as far as endpoints and progression are concerned. Looking at the plot, we can observe that there is a reasonable degree of similarity; in light of how much less expensive the simulation with COMet-Path was with respect to the RMSD Path with parallel tempering or funnel restraints, this result is very interesting. Our simulation was a 1 $\mu s$ run with a single replica, whereas the RMSD Path was converged using five replicas running for 1 $\mu$ each . The difference between the two

approaches can be quantified by estimating the value of the binding free energy of Dasatinib. The result for the unbinding energy is between 28 and 29 kJ/mol using the RMSD Path, as compared to 25 kJ/mol using COMet-Path. The small discrepancy is due to the tight restraint on the distance from the path used in the RMSD Path simulations (More details can be found in the Supporting Information).

The COMet-Path optimized PCV free energy reweighted onto the RMSD PCVs (Fig. 3 panel F) allows for a direct visual comparison of the free energy surface with respect to the original RMSD PCVS result (Fig. 3 panel E). To facilitate the juxtaposition between the two free energy surfaces, we imposed a cutoff on the conformations in the reweighted map (panel F) to mimic the funnel-shaped restraining potential used in the work of Ref. 28. As before, the two surfaces are very similar, the main differences being limited to unbound states at higher values of $s$. However, these can be explained by the absence of strict constraints on the COMet-Path $z$ variables, which allowed for the exploration of numerous conformations far from the reference snapshots, which were then cut off when imposing the funnel-restraint mimic. This limits the sampling of the unbound region in our reweighted map and increases its free energy value considerably. Nevertheless, the general shape of the map is well conserved, especially closer to minimum free energy basins, where most of the sampling is concentrated.

**Comparison with other methods**. It is also interesting to compare the performance of our method with alternative approaches. A widely used path-based approach for this kind of simulations is the Finite Temperature String method.[20] In a way similar to other path-based methods, it works with a set of states that define the transition path. These are then iteratively optimized starting from an initial guess, based on the drift observed during short restrained MD runs. We have applied this method to try and refine both paths 3 and 4. The results can be seen in Fig. 5 A and B. In the case of path 3, the optimization proceeds smoothly, at variance from path 4, which was chosen to be far from the optimal low-free-energy path. As expected, the local optimization used by the String method causes

it to fail to find the faraway optimal path. On the other hand, Path CVs can cope with it due to the inclusion of the $z$ variable, which allows for nonlocal exploration. This advantage is compounded by the use of Metadynamics.

Of great interest is also the comparison with SGOOP.[37] Our method is inspired by it and shares many similarities; Trialanine was used as a test system for both methods. As expected, SGOOP works very well on trialanine, as it can be seen in Fig. 5 C and D. The free energy surface is almost identical to the one obtained using COMet-Path or simple Metadynamics. However, simple CVs or their combination will struggle with complex, winding paths, as it is the case in many ligand binding or conformational changes studies. In these cases, the important variables for the various transitions might change along the path, favoring the use of path-based approaches such as COMet-Path. To show this, we have divided our Dasatinib / c-Src path into two halves. The spectral gap optimization results show that different sets of variables would describe the two sections appropriately (SI Fig. 1). A path with a sufficiently wide metric will work much better in these cases since it would guide the simulation to explore the whole transition. Furthermore, in very complex cases COMet-Path could devise a changing metric along the path, where a global metric is not optimal for all its regions.

# 3 Conclusions

We have presented here an efficient and automatic method for selecting an optimal metric for Path Collective Variables. The method, named COMet-Path, is inspired by the spectral gap maximization approach developed by Tiwary et al.;[37] it combines simple collective variables to bypass the computational cost and alignment issues arising from the use of RMSD as a metric and to possibly extend the applicability of the PCVs to more complex biological processes, difficult to capture with RMSD only.

We have successfully tested this method on two systems. The first, a simple trialanine

peptide in vacuo, was chosen as a toy model for which ideal CVs, three $\Phi$ angles, are known. Testing a number of paths interconnecting different minimum free energy basins, we observed that COMet-Path correctly identifies the appropriate angle CVs, by increasing their weight coefficients according to their respective significance in driving the dynamics. What is more, its non-local exploration properties (shared with the original PathCV) allows it to find an optimal low-free-energy path even when starting from a sub-optimal path.

When applying our method to the realistic and significantly more challenging case of Dasatinib binding to the c-Src kinase, we were able to efficiently achieve results comparable to those reported in the literature, without the need to employ computationally expensive techniques such as multiple replica parallel tempering Metadynamics. A single replica simulation and an harmonic wall on the $z$ variable to improve convergence were employed, and we observed a notable reduction of computational time needed to achieve a reasonable free energy convergence.

These results show indeed that COMet-Path is a significant improvement to the RMSD metric typically used with PCvs. The clear advantage with respect to the original RMSD (or DRMSD) PCV implementation is that by including all the order parameters necessary for a complete description of the path, the algorithm can clearly distinguish the value of different paths and different metastable states on-path, revealing more details of the underlying mechanisms. Moreover, once the combination of CVs to be used for the metric in a specific system is clear, the same optimal path can be re-used to compute the free energy landscape associated with similar systems; for instance, it could be used for other ligands binding to the same protein or for targets harboring different mutations. Furthermore, the choice of CVs that can be included in COMet-Path is not limited and the coefficients can be changed along the path. As an example, variables describing the degree of protein or ligand solvation or similar complex and non-geometrical properties, that would be otherwise impossible to capture with a simple RMSD metric, could be chosen and their relative coefficients might be increased where needed. We believe that our method is thus suitable for many systems

of biological and pharmacological interest characterized by rough free energy minima and amenable to a path description.

## Acknowledgement

FLG acknowledges EPSRC [grant no EP/P022138/1; EP/P011306/1; EP/M013898/1] for financial support. HecBioSim [EPSRC grant no EP/P022138/1], Archer, JADE, the Hartree Centre, the Barcelona Supercomputing Center, and PRACE are acknowledged for computer time. The authors acknowledge Giorgio Saladino for helpful discussions.

## Supporting Information Available

A Full Derivation of the Transition Matrix can be found in the Supporting Information, together with the values of the optimized COMet-Path CVs for trialanine using alternative settings, for Dasatinib binding to the c-Src. Plots showing the variation of the optimized values along the path, detailing the convergence of the simulations and a comparison of the free energies of the COMet-Path and RMSD PCVs results as a function of the unbound limit cutoff can also be found. The Supporting Information is available free of charge on the ACS Publications website. This material is available free of charge via the Internet at `http://pubs.acs.org/`.

## References

(1) Shaw, B. D. E.; Deneroff, M. M. et al. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Commun. ACM* **2008**, *51*, 91–97.

(2) Harvey, M. J.; Giupponi, G. et al. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* **2009**, *5*, 1632–1639.

(3) Stone, J. E.; Phillips, J. C. et al. Accelerating Molecular Modeling Applications with Graphics Processors. *J. Comput. Chem.* **2007**, *28*, 2918–2640.

(4) Pronk, S.; Páll, S. et al. GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845–854.

(5) Lindorff-Larsen, K.; Piana, S. et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950–1958.

(6) Shaw, D. E.; Maragakis, P. et al. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341–346.

(7) Robertson, M. J.; Tirado-Rives, J. et al. Improved Peptide and Protein Torsional Energetics with the OPLS-AA Force Field. *J. Chem. Theory Comput.* **2015**, *11*, 3499–3509.

(8) Pietrucci, F. Strategies for the exploration of free energy landscapes: Unity in diversity and challenges ahead. *Rev. Phys.* **2017**, *2*, 32–45.

(9) Saladino, G.; Estarellas, C. et al. Recent Progress in Free Energy Methods. In *Comprehensive Medicinal Chemistry III*; Elsevier, 2017; pp 34–50.

(10) Bolhuis, P. G.; Chandler, D. et al. Transition Path Sampling : Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

(11) Faradjian, A. K.; Elber, R. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* **2004**, *120*, 10880–10889.

(12) Patey, G. N.; Valleau, J. P. A Monte Carlo method for obtaining the interionic potential of mean force in ionic solution. *J. Chem. Phys.* **1975**, *63*, 2334.

(13) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.

(14) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562–12566.

(15) Zhou, R. Replica exchange molecular dynamics method for protein folding simulation. *Methods Mol. Biol.* **2007**, *350*, 205–223.

(16) Isralewitz, B.; Gao, M. et al. Steered molecular dynamics and mechanical functions of proteins. *Curr. Opin. Struct. Biol.* **2001**, *11*, 224–230.

(17) Sutto, L.; D'Abramo, M. et al. Comparing the efficiency of biased and unbiased molecular dynamics in reconstructing the free energy landscape of Met-enkephalin. *J. Chem. Theory Comput.* **2010**, *6*, 3640–3646.

(18) Wales, D. J. Discrete path sampling. *Mol. Phys.* **2002**, *100*, 3285–3305.

(19) Mills, G.; Jonsson, H. et al. Reversible Work Transiton State Theory: Application to Dissociative Adsorption of Hydrogen. *Surf. Sci.* **1995**, *324*, 305–337.

(20) E, W.; Ren, W. et al. Finite temperature string method for the study of rare events. *J. Phys. Chem. B* **2005**, *109*, 6688–6693.

(21) Sutto, L.; Marsili, S. et al. New advances in metadynamics. *WIREs Comput Mol Sci* **2012**, *2*, 771–779.

(22) Cavalli, A.; Spitaleri, A. et al. Investigating drug-target association and dissociation mechanisms using metadynamics-based algorithms. *Acc. Chem. Res.* **2015**, *48*, 277–285.

(23) Branduardi, D.; Gervasio, F. L. et al. From A to B in free energy space. *J. Chem. Phys.* **2007**, *126*, 054103.

(24) Barducci, A.; Bussi, G. et al. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008**, *100*, 1–4.

(25) Díaz Leines, G.; Ensing, B. Path finding on high-dimensional free energy landscapes. *Phys. Rev. Lett.* **2012**, *109*, 5–8.

(26) Fidelak, J.; Juraszek, J. et al. Free-Energy-Based Methods for Binding Profile Determination in a Congeneric Series of CDK2 Inhibitors. *J. Phys. Chem. B* **2010**, *114*, 9516–9524.

(27) Saladino, G.; Gauthier, L. et al. Assessing the performance of metadynamics and path variables in predicting the binding free energies of p38 inhibitors. *J. Chem. Theory Comput.* **2012**, *8*, 1165–1170.

(28) Morando, M. A.; Saladino, G. et al. Conformational Selection and Induced Fit Mechanisms in the Binding of an Anticancer Drug to the c-Src Kinase. *Sci. Rep.* **2016**, *6*, 24439.

(29) Provasi, D.; Artacho, M. C. et al. Ligand-Induced modulation of the Free-Energy landscape of G protein-coupled receptors explored by adaptive biasing techniques. *PLoS Comput. Biol.* **2011**, *7*.

(30) Limongelli, V.; Bonomi, M. et al. Molecular basis of cyclooxygenase enzymes (COXs) selective inhibition. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 5411–6.

(31) Bešker, N.; Gervasio, F. L. Using Metadynamics and Path Collective Variables to Study Ligand Binding and Induced Conformational Transitions. In *Computational Drug Discovery and Design*; Springer New York, 2012; pp 501–513.

(32) Juraszek, J.; Saladino, G. et al. Efficient numerical reconstruction of protein folding kinetics with partial path sampling and pathlike variables. *Phys. Rev. Lett.* **2013**, *110*, 1–5.

(33) Kearsley, S. K. On the orthogonal transformation used for structural comparisons. *Acta Cryst.* **1989**, *45*, 208–210.

(34) Bonomi, M.; Branduardi, D. et al. The unfolded ensemble and folding mechanism of the C-terminal GB1 $\beta$-hairpin. *J. Am. Chem. Soc.* **2008**, *130*, 13938–13944.

(35) Tribello, G. A.; Bonomi, M. et al. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2013**, *185*, 604–613.

(36) Pietropaolo, A.; Branduardi, D. et al. A Chirality-Based Metrics for Free-Energy Calculations in Biomolecular Systems. *J. Comput. Chem.* **2011**, *32*, 2627–2637.

(37) Tiwary, P.; Berne, B. J. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 2839–2844.

(38) Berezhkovskii, A.; Szabo, A. Time scale separation leads to position-dependent diffusion along a slow coordinate. *J. Chem. Phys.* **2011**, *135*, 1–5.

(39) Pérez-Hernández, G.; Paul, F. et al. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*.

(40) Dixit, P. D.; Jain, A. et al. Inferring transition rates of networks from populations in continuous-time Markov processes. *J. Chem. Theory Comput.* **2015**, *11*, 5464–5472.

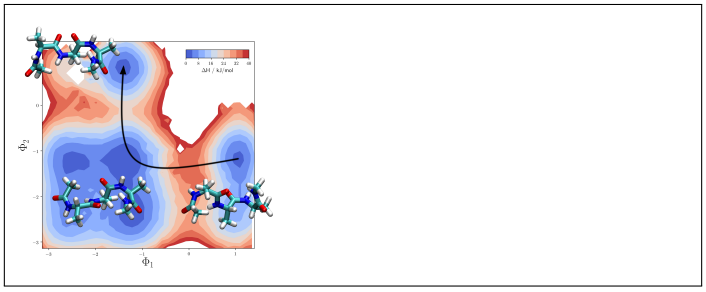
(41) Tiwary, P.; Parrinello, M. A Time-Independent Free Energy Estimator for Metadynamics. *J. Phys. Chem. B* **2015**, *119*, 736–742.

(42) Dickson, B. M.; Huang, H. et al. Unrestrained computation of free energy along a path. *J. Phys. Chem. B* **2012**, *116*, 11046–11055.

(43) Zinovjev, K.; Tuñõn, I. Exploring chemical reactivity of complex systems with path-based coordinates: Role of the distance metric. *J. Comput. Chem.* **2014**, *35*, 1672–1681.

# 4 Figures

# Graphical TOC Entry



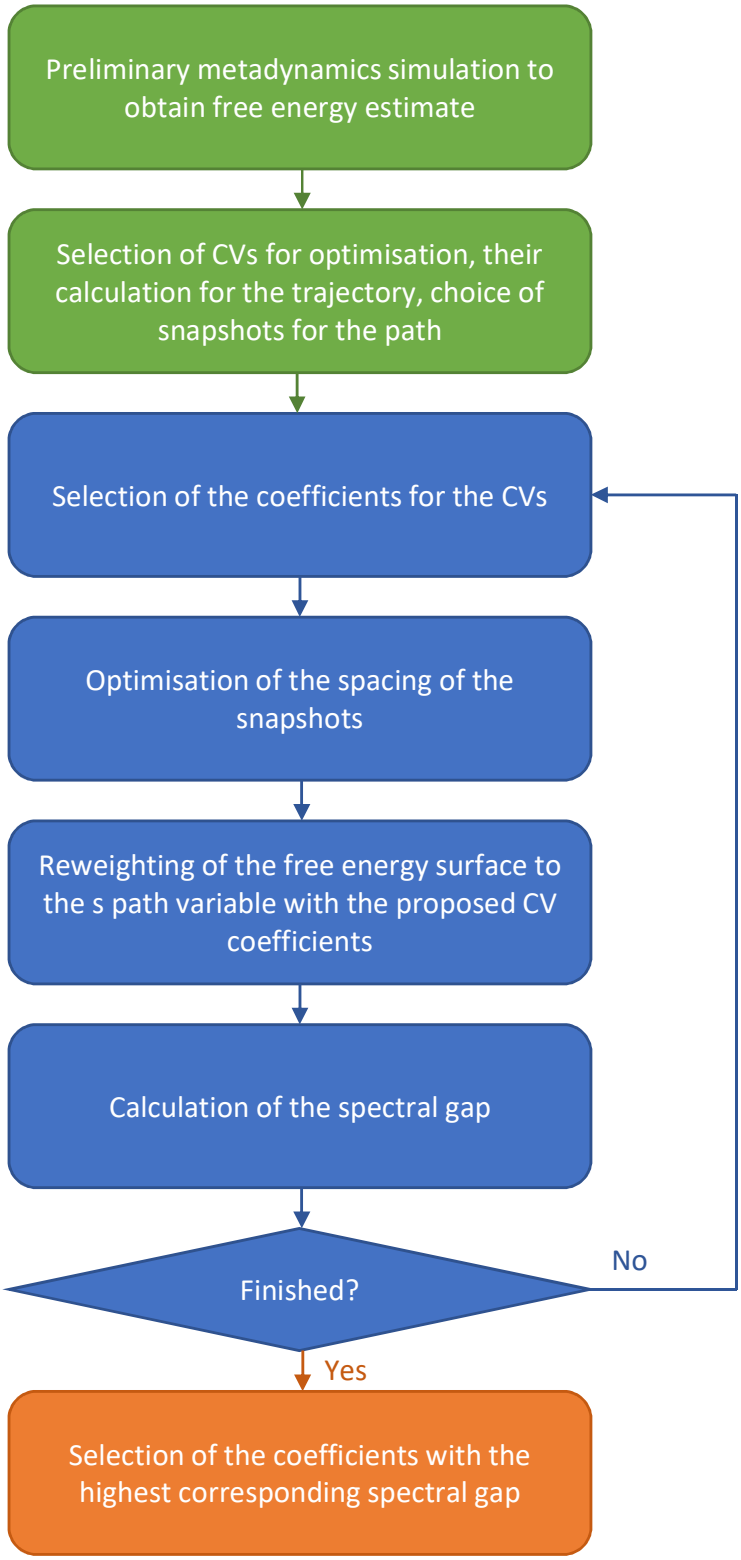



Figure 1: Schematic representation of the typical COMet-Path workflow. After an initial exploration of the FES with MetaD simulations (in orange), the optimisation of the coefficients is performed iteratively in postprocessing (in blue).

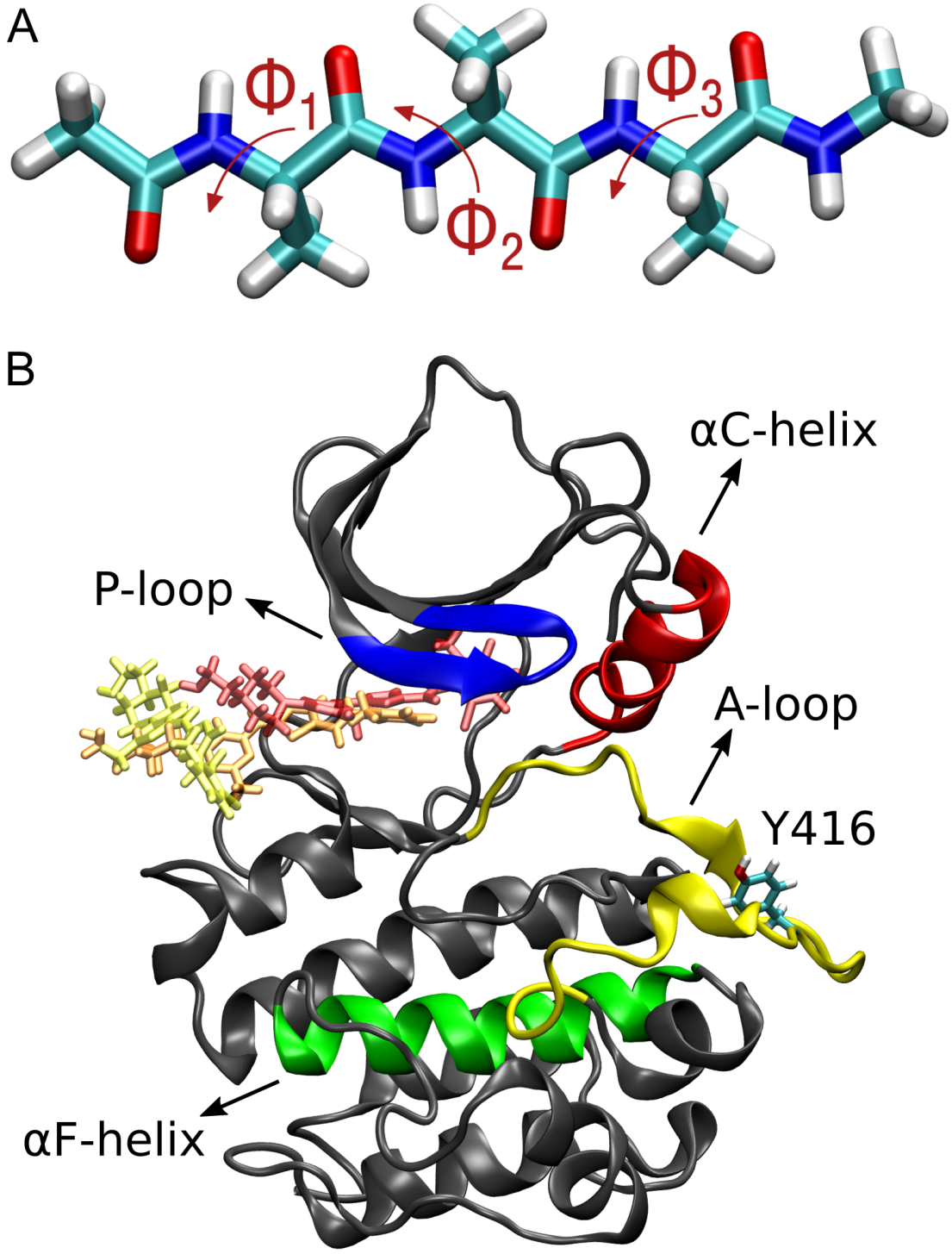
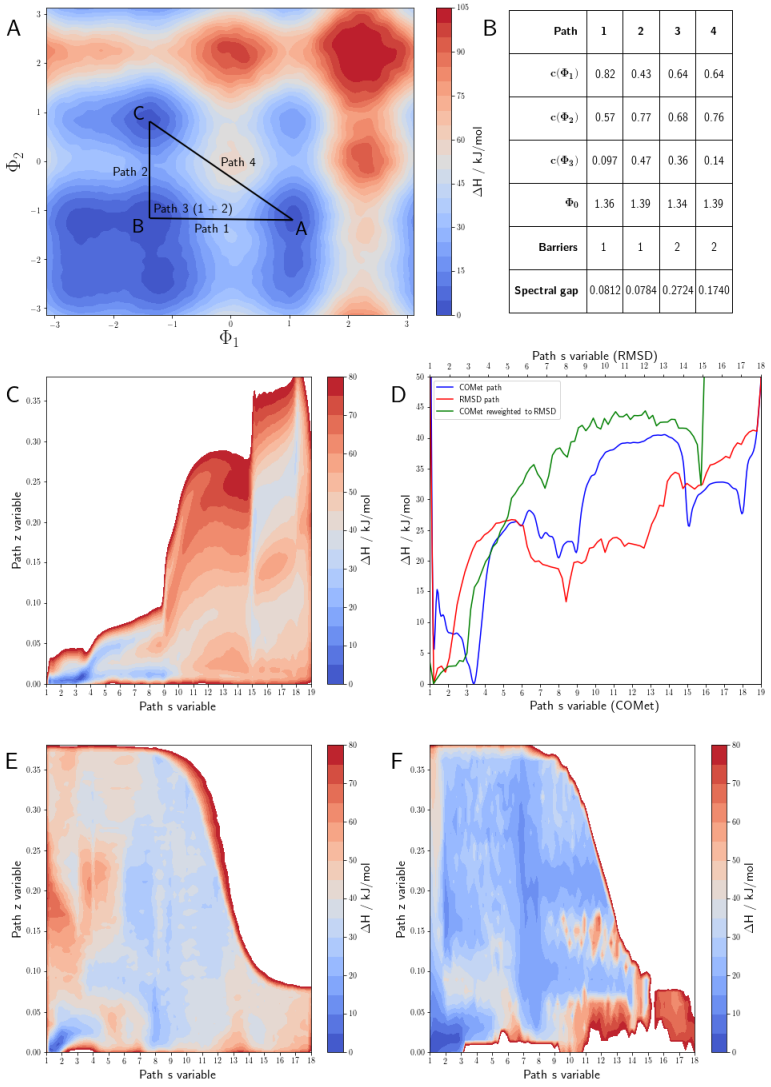



Figure 2: A. A trialanine peptide with the Φ angles labeled explicitly. B. Dasatinib binding to a c-Src kinase. Three representative structures of the unbinding path are shown for the ligand (from red to yellow).




| Path | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $c(\Phi_1)$ | 0.82 | 0.43 | 0.64 | 0.64 |
| $c(\Phi_2)$ | 0.57 | 0.77 | 0.68 | 0.76 |
| $c(\Phi_3)$ | 0.097 | 0.47 | 0.36 | 0.14 |
| $\Phi_0$ | 1.36 | 1.39 | 1.34 | 1.39 |
| **Barriers** | 1 | 1 | 2 | 2 |
| **Spectral gap** | 0.0812 | 0.0784 | 0.2724 | 0.1740 |

Figure 3: A. The four proposed paths (in black) on the 2D projection of the free energy surface of trialanine. B. A table summarizing the optimized coefficients for the paths shown in panel A and the corresponding numbers of barriers and spectral gaps values. Lower panels: The free energy surface of the Dasatinib / c-Src system as obtained using COMet-Path (panel C), using RMSD path with funnel-like restraints (E) and the same free energy obtained with COMet-Path and reweighted against the RMSD path with funnel-like restraints (F). In panel D, the 1D projections of these three free energies are compared in blue, red and green respectively).

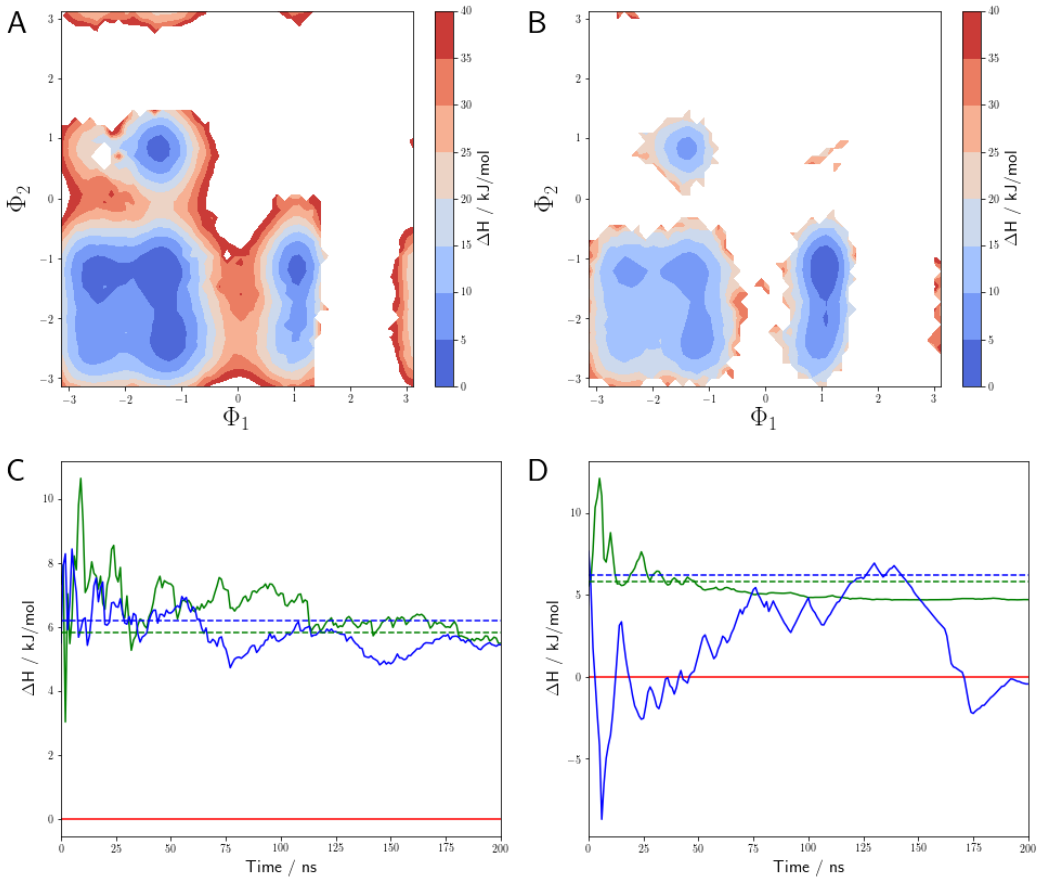



Figure 4: A. The reweighted free energy surface obtained from a COMet-Path simulation on path 3. B. The equivalent reweighted free energy surface obtained using RMSD path instead. The time evolution of the free energy estimates for the three minima is shown in panels C and D for the COMet-Path and RMSD path simulations respectively. The energies are shown relative to the central minimum (labeled B in Fig. 3 A and shown as a red line here). The green solid line corresponds to the minimum A, while a blue solid line is used for the minimum C. Dashed lines of the same colors show reference values obtained directly from the 2D Metadynamics on the two angles.

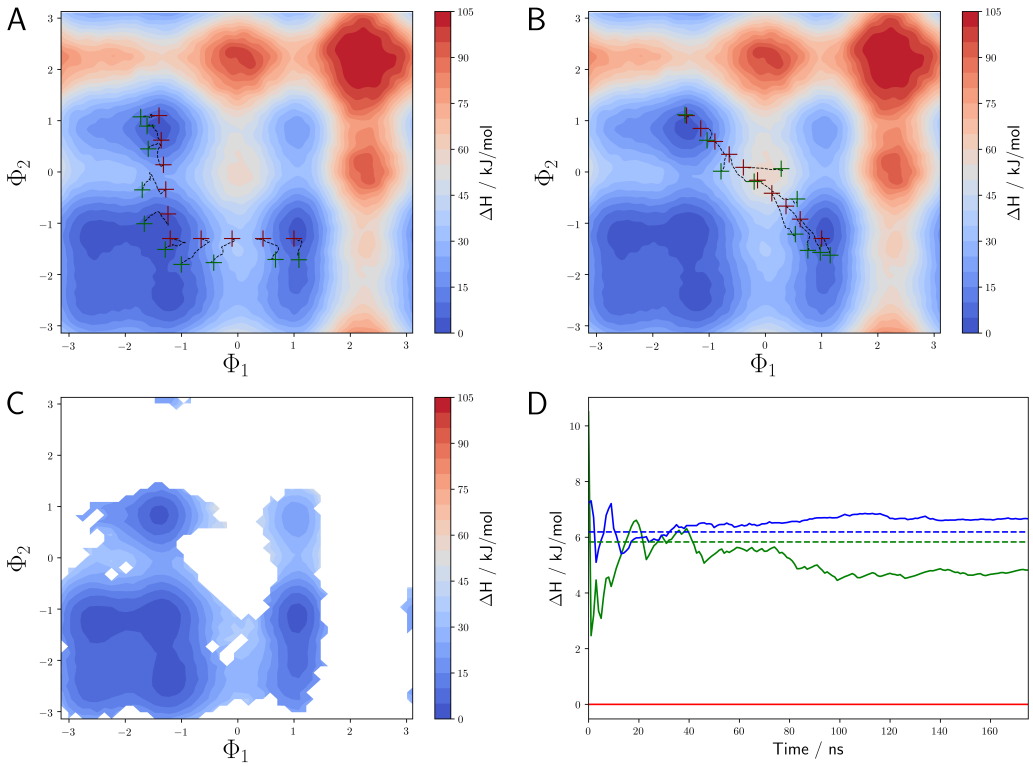



Figure 5: A. The evolution of the optimal path using the Finite Temperature String method, starting from path 3. All six torsional angles were optimized. Red marks correspond to the starting points, green marks correspond to the endpoints. The underlying free energy projection is the same as in Fig. 3. B. The equivalent evolution of the optimal path starting from path 4 instead. C. The projection of the free energy surface of trialanine obtained by reweighting the results of a SGOOP simulation. D. The free energy estimates for the different minima over the course of the SGOOP simulation, presented in the same way as in Fig. 4.